

A vertical strip on the left side of the slide contains two grayscale micrographs of pollen grains. The top image shows a highly textured, spherical grain with a complex surface pattern. The bottom image shows a smoother, more rounded grain.

# **An automated image analysis platform for palynological specimens**

---

**Surangi W. Punyasena**

Department of Plant Biology  
University of Illinois

**Kenton McHenry**

National Center for Supercomputing Applications  
University of Illinois

Fossil pollen =  
a microscopic census of  
past vegetation,  
preserved in  
geologic sediments

# **POLLEN APPLICATIONS**

---

Biostratigraphy

Paleoclimate

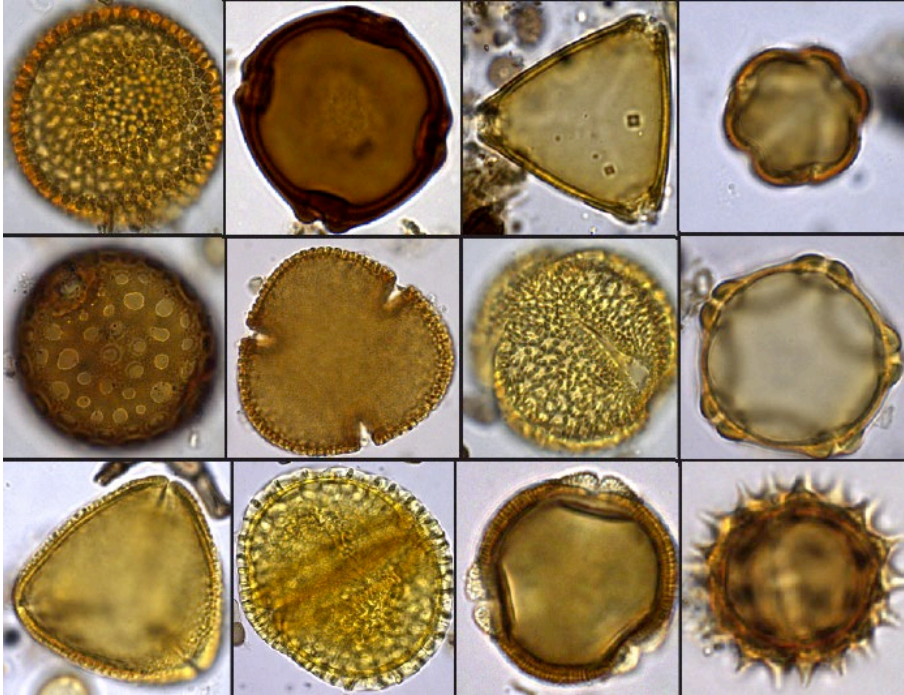
Paleoecology

Plant evolution

Forensics

# POLLEN AS “BIG DATA”

---



470 million years of plant history

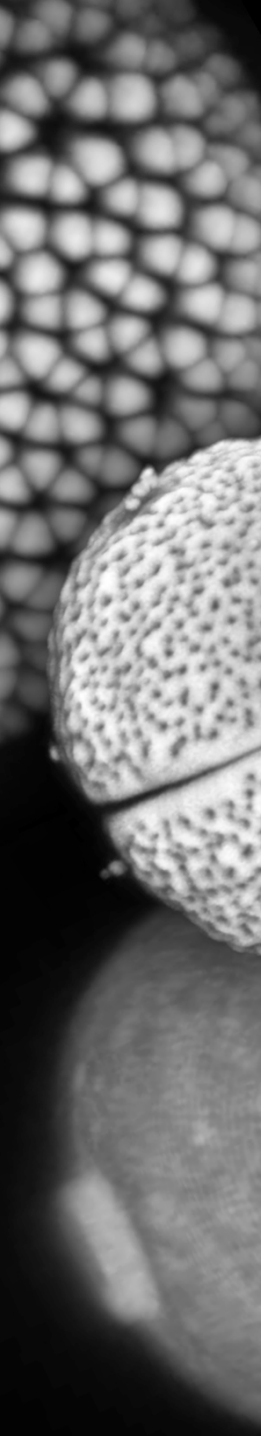
Billions of potential specimens

Continuous deposition across a range of environments

# WHAT HOLDS US BACK?

---

- Limited data acquisition rates
  - Highly specialized skill
  - Counting methods unchanged for ~100 years
- No formal mechanisms for evaluating identification accuracy and consistency
  - Dependent on qualitative descriptions
  - No estimates for identification confidence
  - Uncertainty cannot be propagated
- Variable taxonomic precision
  - “Lowest common denominator” identifications
  - Species identifications often rare or absent
  - Limits ecological and evolutionary interpretation



# WHAT HOLDS US BACK?

---

- Limited data acquisition rates
  - Highly specialized skill
  - Counting methods unchanged for ~100 years
- No formal mechanisms for evaluating identification accuracy and consistency
  - Dependent on qualitative descriptions
  - No estimates for identification confidence
  - Uncertainty cannot be propagated
- Variable taxonomic precision
  - “Lowest common denominator” identifications
  - Species identifications often rare or absent
  - Limits ecological and evolutionary interpretation

# WHAT HOLDS US BACK?

---

- Limited data acquisition rates
  - Highly specialized skill
  - Counting methods unchanged for ~100 years
- No formal mechanisms for evaluating identification accuracy and consistency
  - Dependent on qualitative descriptions
  - No estimates for identification confidence
  - Uncertainty cannot be propagated
- Variable taxonomic precision
  - “Lowest common denominator” identifications
  - Species identifications often rare or absent
  - Limits ecological and evolutionary interpretation

## **DATA QUANTITY**

How do we transform pollen analysis into a higher-throughput “big data” discipline?

## **DATA REPRODUCIBILITY**

How do we improve the consistency and accuracy of pollen identifications?

## **TAXONOMIC RESOLUTION**

How do we discriminate among morphologically similar taxa?



## **DATA QUANTITY**

How do we transform pollen analysis into a higher-throughput “big data” discipline?

## **DATA REPRODUCIBILITY**

How do we improve the consistency and accuracy of pollen identifications?

## **TAXONOMIC RESOLUTION**

How do we discriminate among morphologically similar taxa?

## **DATA QUANTITY**

How do we transform pollen analysis into a higher-throughput “big data” discipline?

## **DATA REPRODUCIBILITY**

How do we improve the consistency and accuracy of pollen identifications?

## **TAXONOMIC RESOLUTION**

How do we discriminate among morphologically similar taxa?

# AUTOMATED POLLEN CLASSIFICATION WORKFLOW

---

Slide imaging

Image segmentation

•Taxonomic classification



Charless Fowlkes  
(UC Irvine)

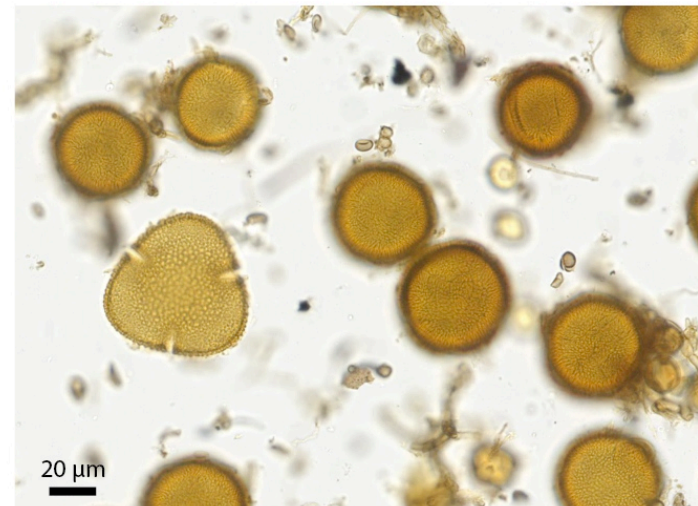
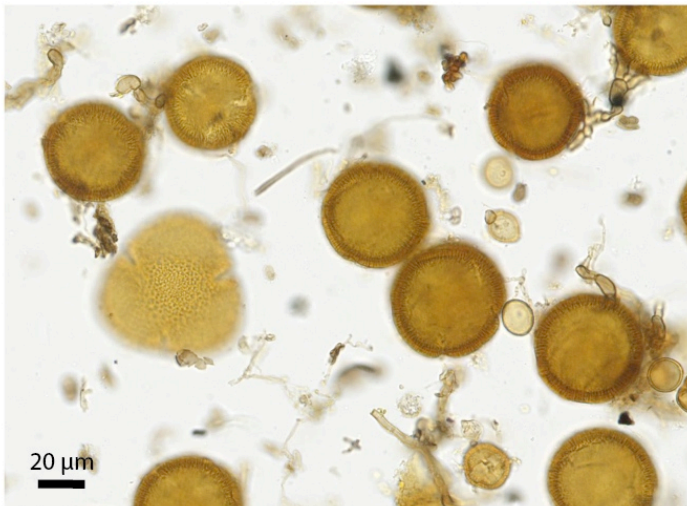
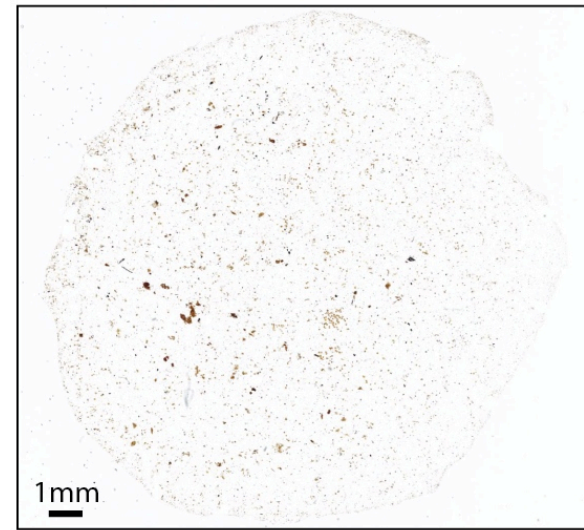
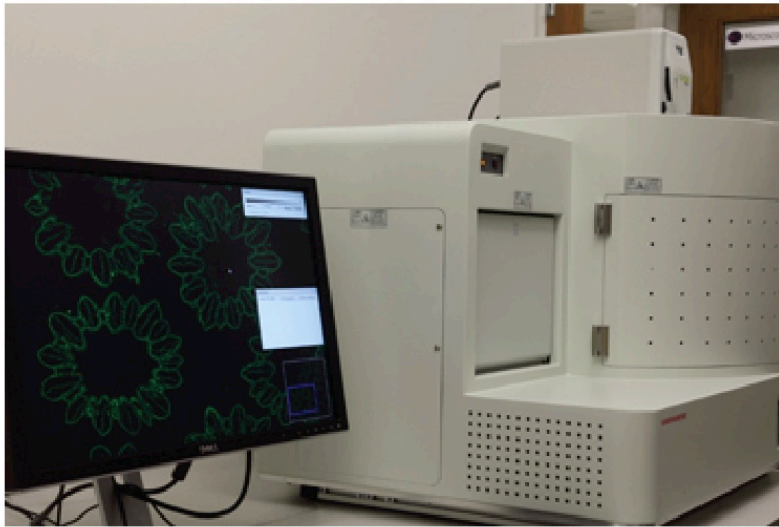


Shu Kong  
(UC Irvine)



Derek Haselhorst  
(U Illinois)

# AUTOMATED IMAGING / VIRTUAL SLIDES



400x, 0.23 μm/pixel

One sample (41 @ 1 μm axial planes) = ~400 GB

# EXPERT ANNOTATIONS

---

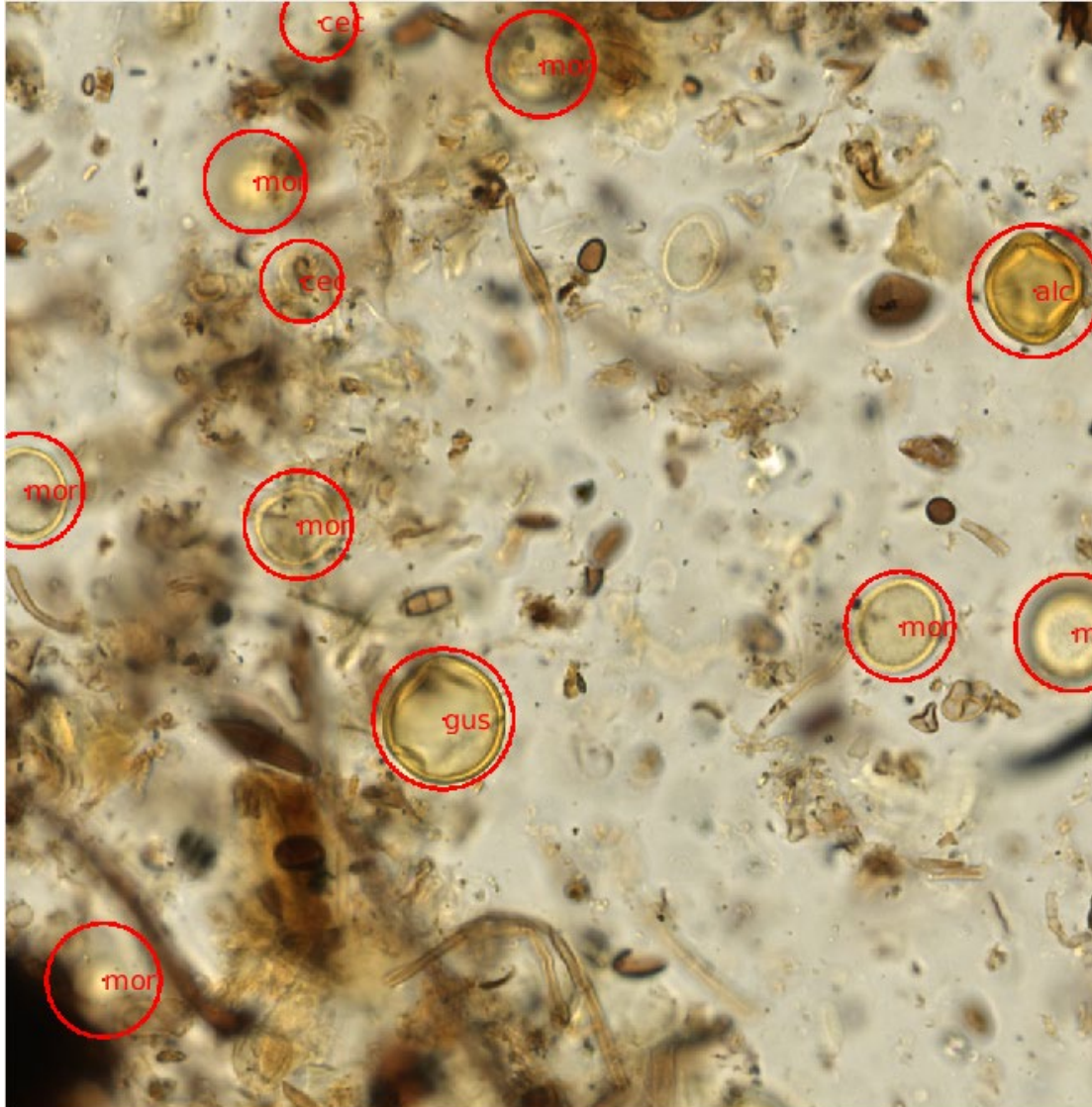
PNG image stacks randomly  
subsampled

Subsample windows viewed with  
virtual microscope by expert

- Metadata tagged for each grain:  
taxon, ID confidence,  
coordinates, radius, slide info

# EXPERT ANNOTATIONS

---



One of 41 planes of view

# SEGMENTATION & CLASSIFICATION

---

Annotated images divided into a training set (70%) and a validation set (30%)

Data augmentation by randomly flipping and rotating images (~2,000 images per pollen type)

Fine-tune pre-trained convolutional neural net (CNN) (Alexnet) (Krizhevsky et al, 2012)

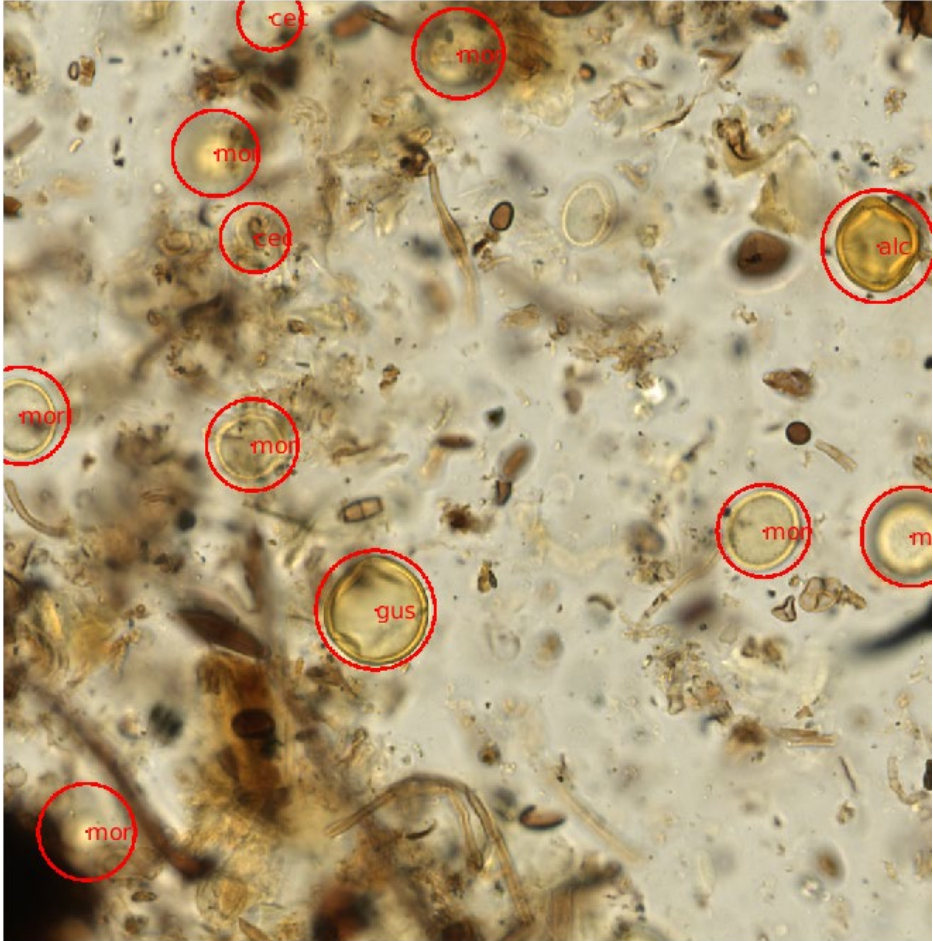
Utilize the radius information to train segmentation masks

Non-maximum suppression used to identify pollen grains

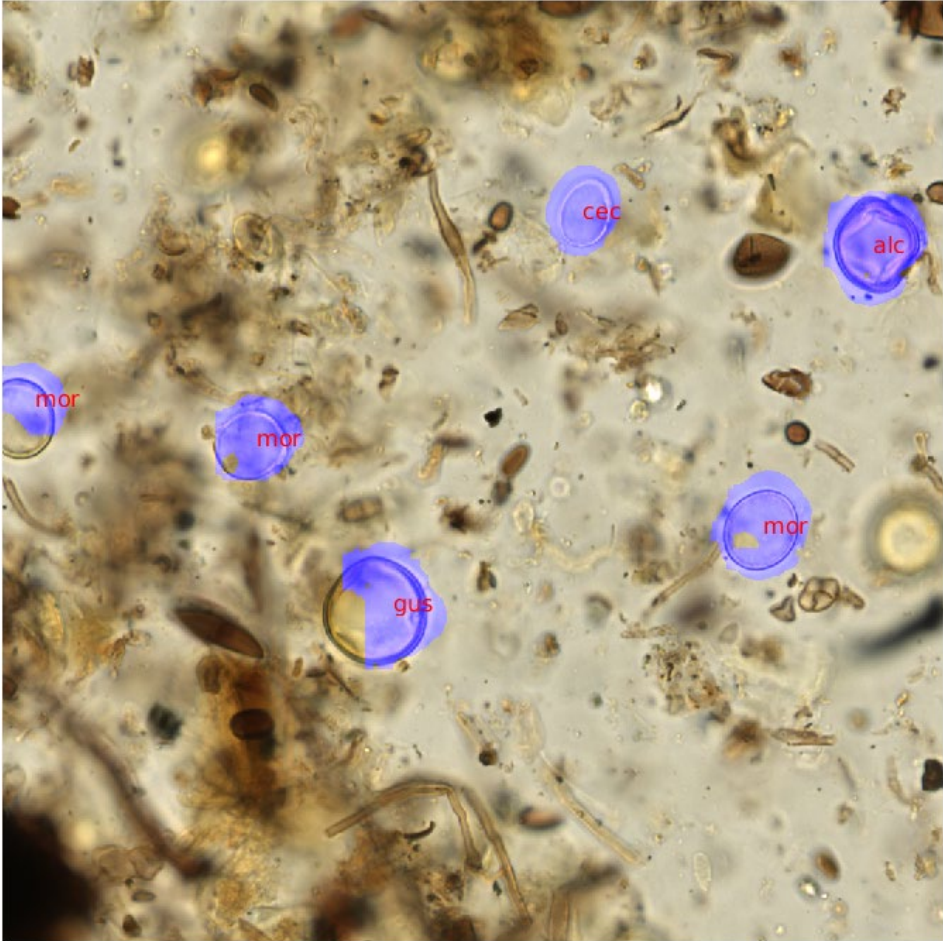


# SEGMENTATION & CLASSIFICATION

---



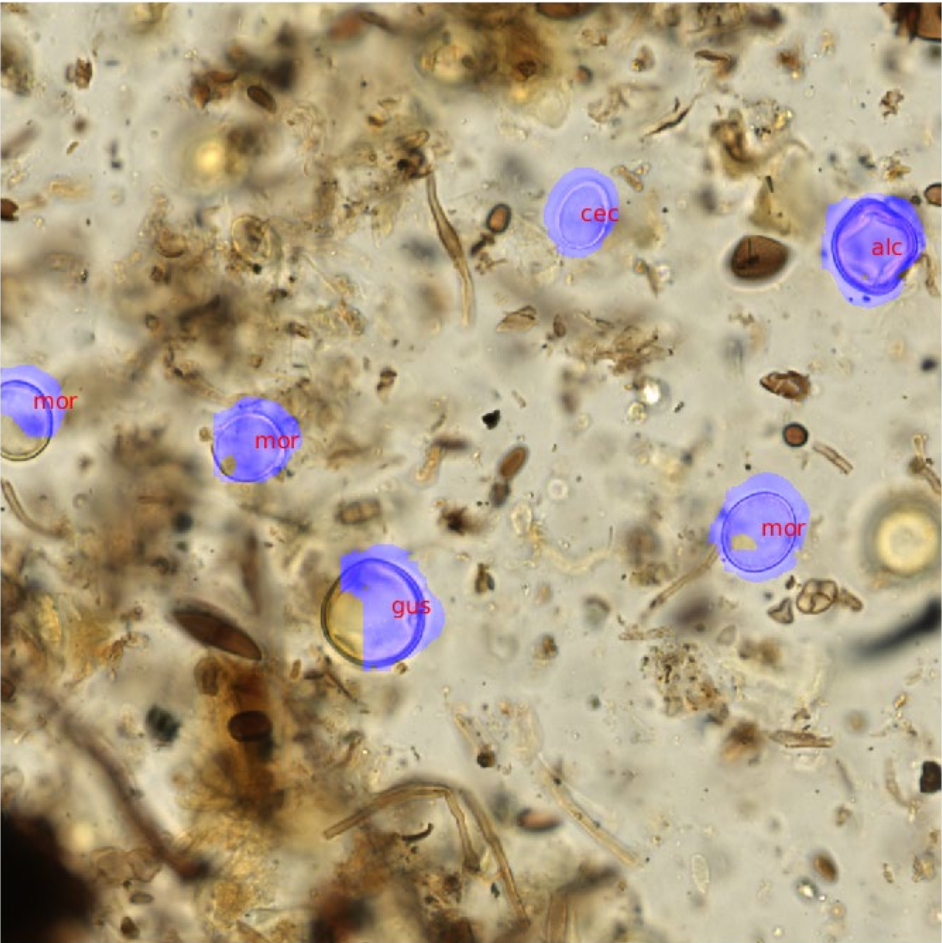
Human



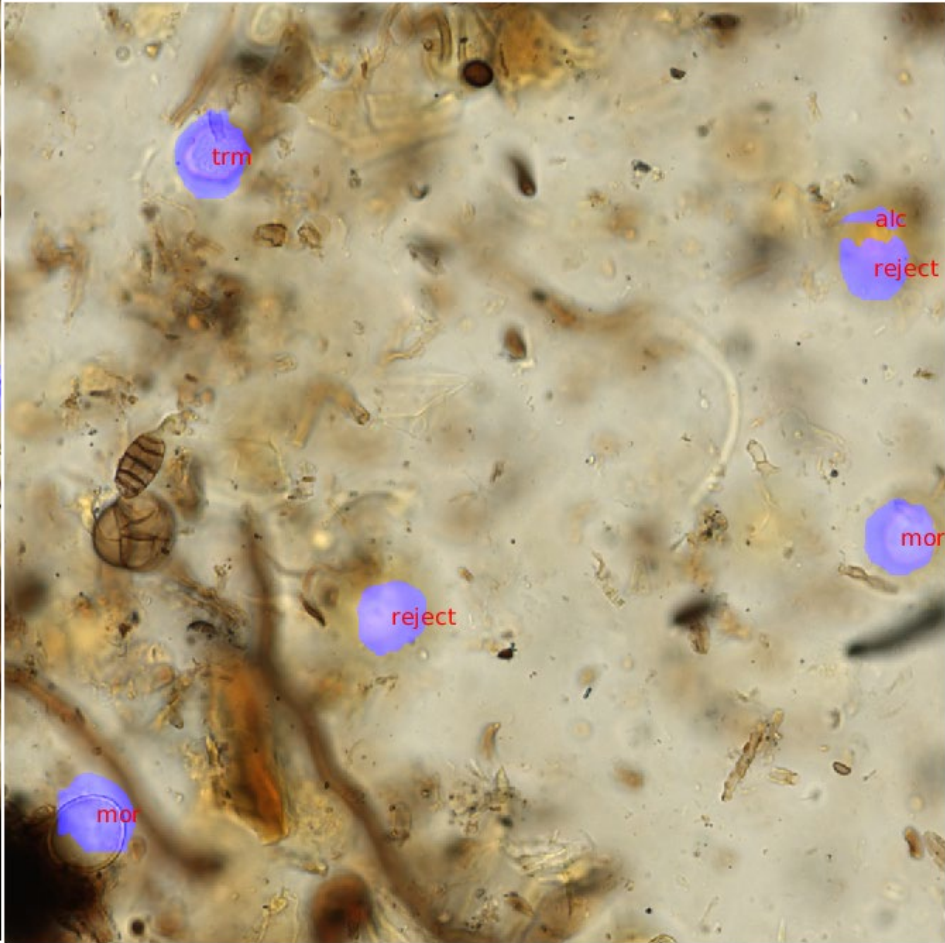
Machine

# SEGMENTATION & CLASSIFICATION

---



**Machine (plane 0)**



**Machine (plane +20)**

# CLASSIFICATION RESULTS

87.25% accuracy for 25 most common/distinctive types

confusion matrix on test set (acc=87.25%)

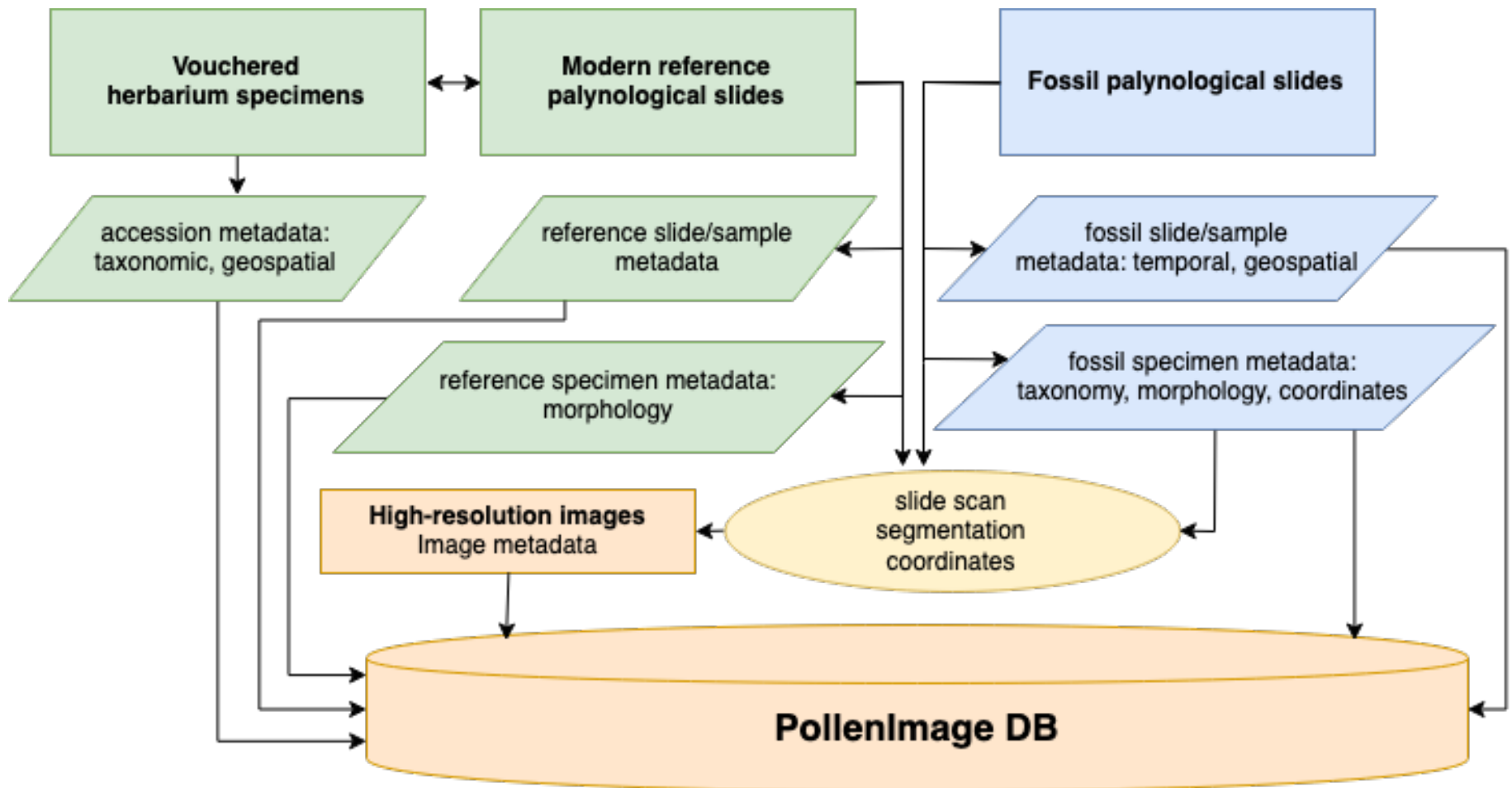
ground-truth label	als	ant	cas	cec	cor	fic	fra	hir	hyr	lae	lue	lyc	mic	mor	oen	ply	qua	sim	slo	tab	tch	unc	vir	vra	zna		
als	0.77	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.19	
ant	0.08	0.62	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08
cas	0.00	0.00	0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.04
cec	0.00	0.00	0.00	0.86	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
cor	0.00	0.00	0.00	0.00	0.89	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02
fic	0.00	0.00	0.00	0.05	0.00	0.71	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
fra	0.00	0.00	0.00	0.00	0.00	0.00	0.94	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
hir	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.79	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
hyr	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
lae	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
lue	0.00	0.00	0.00	0.00	0.07	0.00	0.02	0.00	0.00	0.00	0.91	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
lyc	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
mic	0.00	0.00	0.02	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.83	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
mor	0.00	0.00	0.00	0.10	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.88	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
oen	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.03	0.87	0.06	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ply	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.87	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
qua	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
sim	0.02	0.02	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.88	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05
slo	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.79	0.00	0.00	0.00	0.00	0.00	0.00	0.00
tab	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.94	0.00	0.00	0.00	0.00	0.00	0.00
tch	0.00	0.00	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.80	0.00	0.00	0.00	0.00	0.00
unc	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.83	0.00	0.00	0.00
vir	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.89	0.04	0.00
vra	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.76	0.08
zna	0.06	0.01	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.83

predicted label

the next step

---

# WORKFLOW FOR LARGE-SCALE IMAGING AND ANALYSIS



# PARTNERSHIP WITH KEW AND SMITHSONIAN

---



Alex Antonelli  
(Kew)



Carlos Jaramillo  
(Smithsonian Tropical  
Research Institute)

# PARTNERSHIP WITH KEW AND SMITHSONIAN

---



Alex Antonelli  
(Kew)

## Kew Royal Botanic Gardens

- 40,000 slides/pollen residues
- ~7,000,000 herbarium specimens (unique species and cultivars)

# PARTNERSHIP WITH KEW AND SMITHSONIAN

---

## Smithsonian Tropical Research Institute

- 25,000 species, Alan Graham pollen reference collection (historic slides)
- Smithsonian Pollen Database



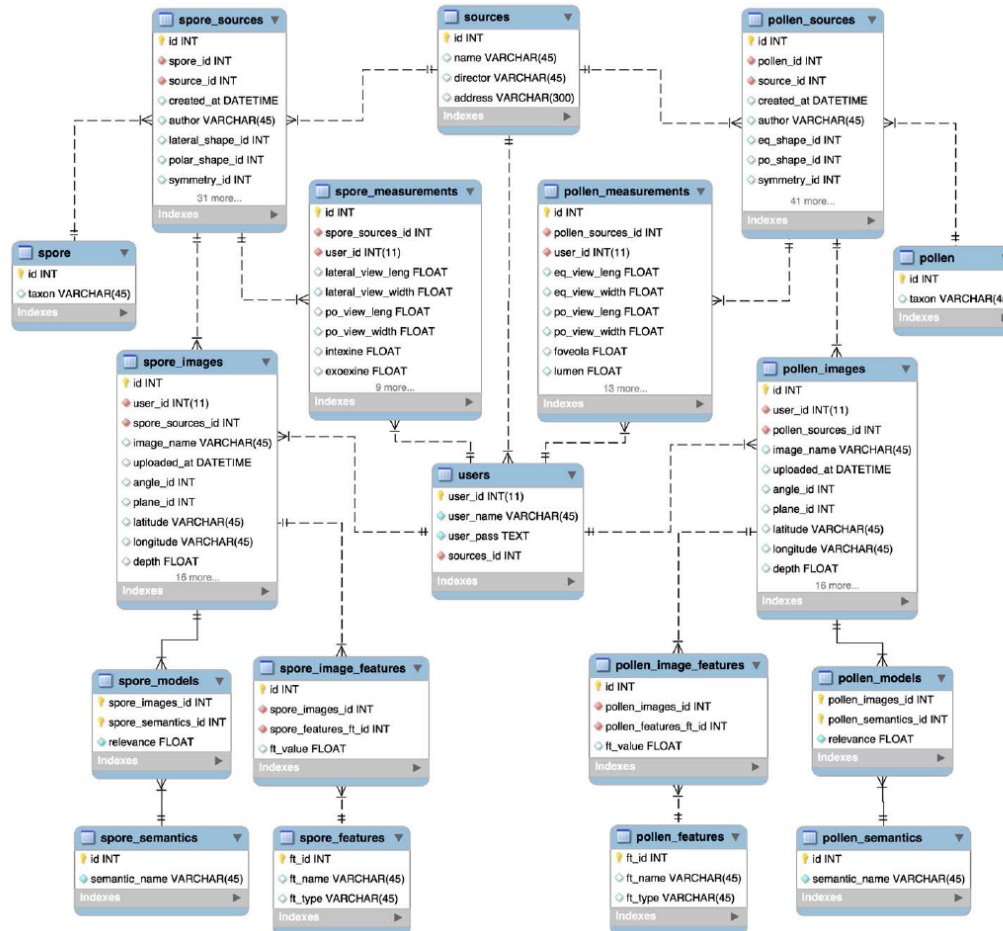
Carlos Jaramillo  
(Smithsonian Tropical  
Research Institute)



# PARTNERSHIP WITH KEW AND SMITHSONIAN

## Smithsonian Pollen Database

Fossil specimens and morphological descriptions



# PARTNERSHIP WITH NCSA

---



Kenton McHenry  
(Principal Research Scientist)



Luigi Marini  
(Senior Programmer)



Rob Kooper  
(Senior Programmer)

# CLOWDER (NCSA)

customizable web interface for data ingestion and analysis

Clowder Explore ▾ Help ▾

Search



Sign up

Login

## Welcome to Clowder

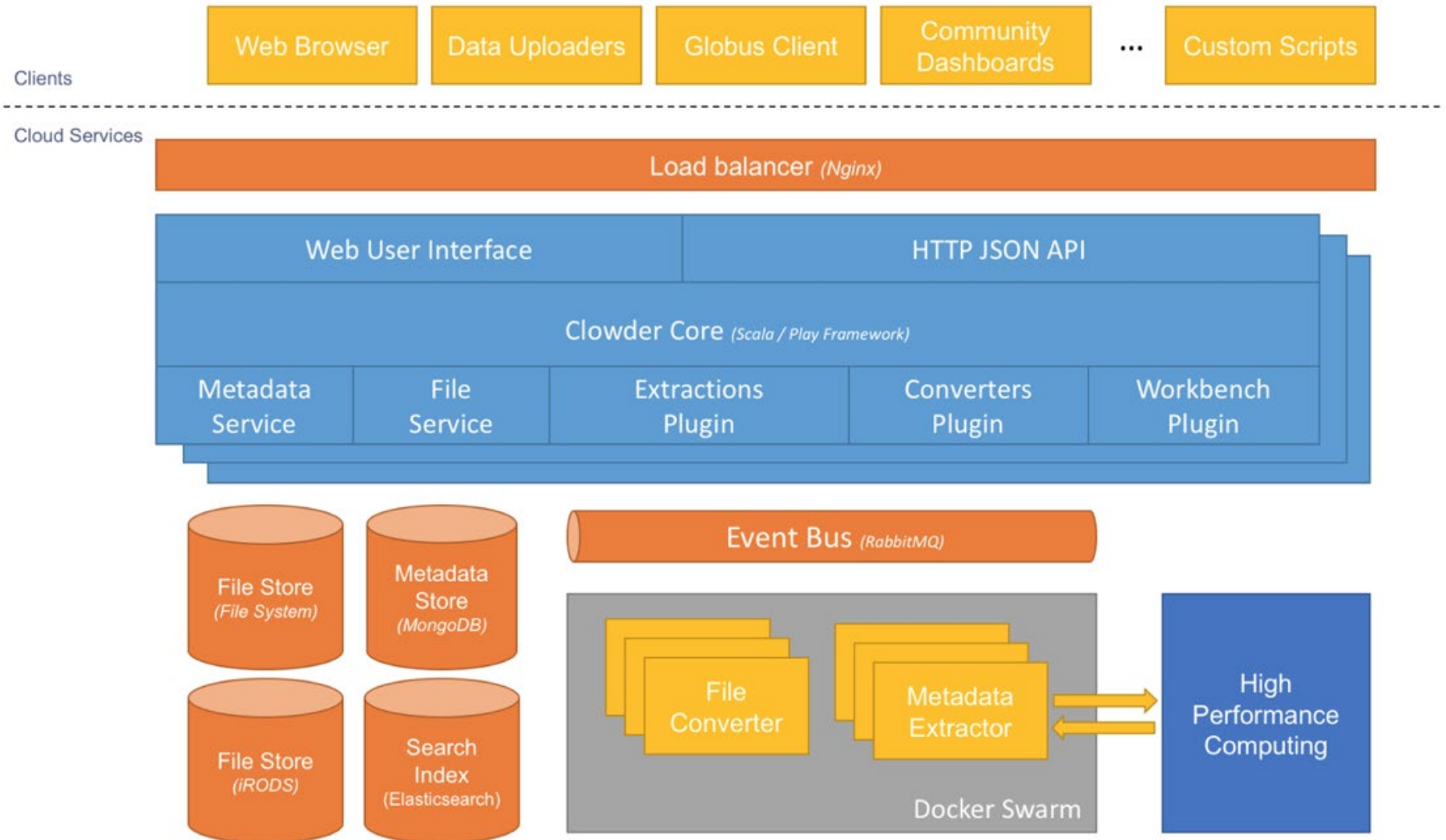
Welcome to Clowder, a scalable data repository where you can share, organize and analyze data.

### Resources

Spaces	0
Collections	0
Datasets	0
Files	0
Bytes	0 B
Users	3

# CLOWDER (NCSA)

customizable web interface for data ingestion and analysis



other workflows

---



Ingrid Romero  
(U Illinois)



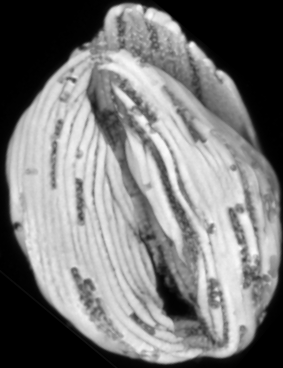
Charless Fowlkes  
(UC Irvine)



Shu Kong  
(UC Irvine)

# Putative Detarioideae legume – *Striatopollis catatumbus*

FOSSIL



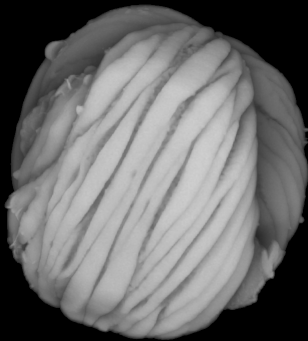
South America



Africa

EXTANT

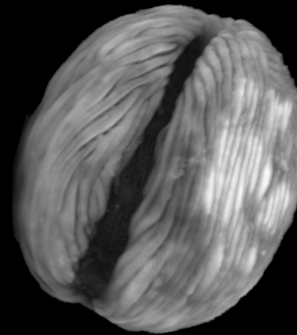
*Macrolobium*



*Crudia*



*Anthonotha*



*Isoberlinia*

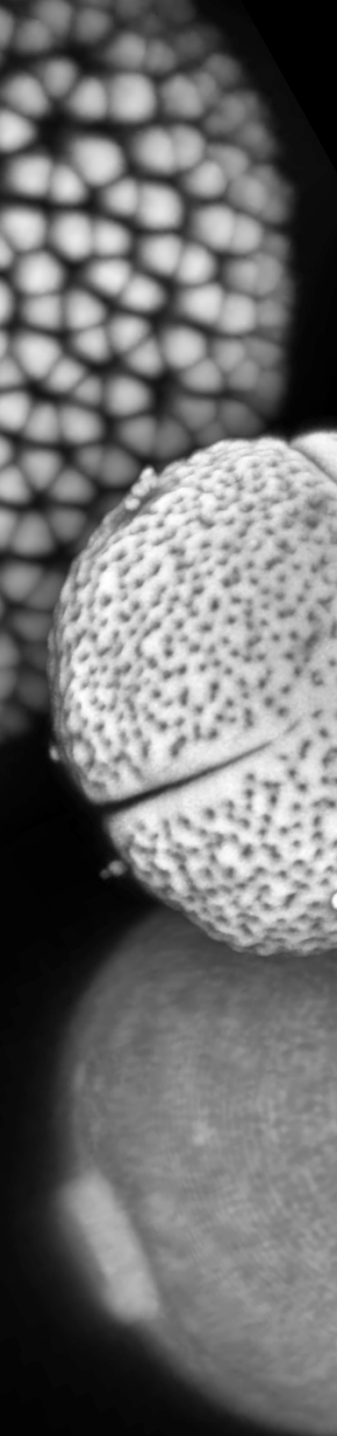


10  $\mu$ m

# CONCLUSIONS

---

- Automation of pollen counts possible
  - Consistent segmentation
  - High accuracy for common, distinct types
  - Model performance poorest on morphologically similar and rare types
- Clowder will allow us to efficiently scale up our analytical pipeline
  - Intuitive interface for non-programmers
  - Real-time analysis and updates
- Accuracy should improve with larger and more balanced image training sets
  - CNNs can generalize from reference images
  - Need high quality images of vouchered, expertly identified specimens!

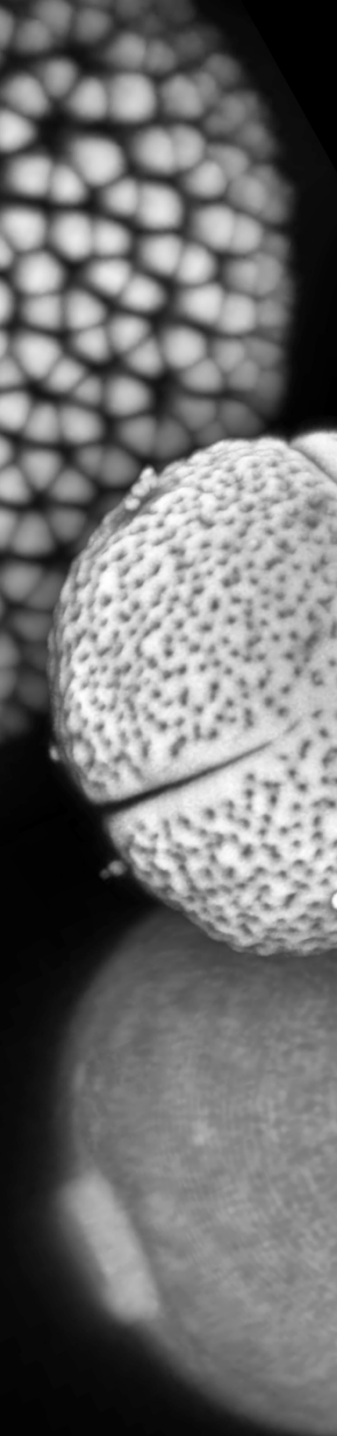




# CONCLUSIONS

---

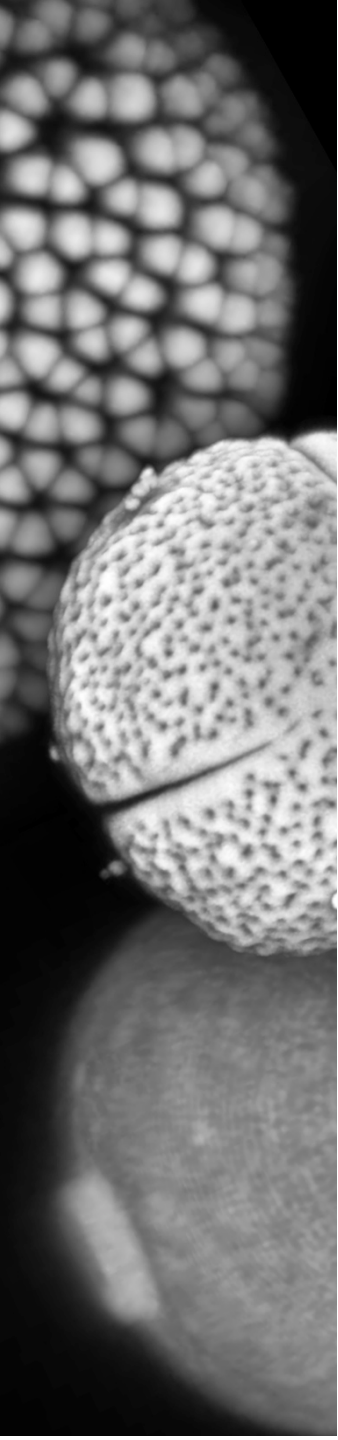
- Automation of pollen counts possible
  - Consistent segmentation
  - High accuracy for common, distinct types
  - Model performance poorest on morphologically similar and rare types
- Clowder will allow us to efficiently scale up our analytical pipeline
  - Intuitive interface for non-programmers
  - Real-time analysis and updates
- Accuracy should improve with larger and more balanced image training sets
  - CNNs can generalize from reference images
  - Need high quality images of vouchered, expertly identified specimens!



# CONCLUSIONS

---

- Automation of pollen counts possible
  - Consistent segmentation
  - High accuracy for common, distinct types
  - Model performance poorest on morphologically similar and rare types
- Clowder will allow us to efficiently scale up our analytical pipeline
  - Intuitive interface for non-programmers
  - Real-time analysis and updates
- Accuracy should improve with larger and more balanced image training sets
  - CNNs can generalize from reference images
  - Need high quality images of vouchered, expertly identified specimens!



Much of the necessary computational infrastructure already exists

Ultimate bottleneck is in  
the availability of  
well-curated, high-resolution images

# COLLABORATORS AND FUNDERS

---

David Tcheng

Illinois Informatics Institute, NCSA

Glenn Fried & Mayandi Sivaguru

Institute for Genomic Biology, UIUC

Enrique Moreno

Smithsonian Tropical Research Institute

---

NSF-DBI – Advances in Biological Informatics

NSF-DBI – Innovations in Biological Imaging & Visualization

NSF-EF – Macrosystems Biology

National Center for Supercomputing Applications (NCSA)

U Illinois Campus Research Board

